Open Science Grid.. An introduction

Ruth Pordes
Fermilab

Introducing myself

- I have been at Fermilab for 25 years (well and 2 years in the "pioneer" '70s).
- I started working on data acquisition for High Energy Physics experiments; worked as a "builder" of the Sloan Digital Sky Survey; managed the DART project for a common data acquisition system for 6 experiments at Fermilab; co-coordinated the CDF/D0 Joint Run II offline projects; co-coordinated the Particle Physics Data Grid SciDAC I collaboratory and introduced Trillium - iVDGL, GridPhyN, PPDG - ad-hoc collaboration from which Open Science Grid derives.
- Now I am variously an Associate Head of the Computing Division at Fermilab, Executive Director of the Open Science Grid, and US CMS Grid Services and Interfaces Coordinator.

Scope of OSG

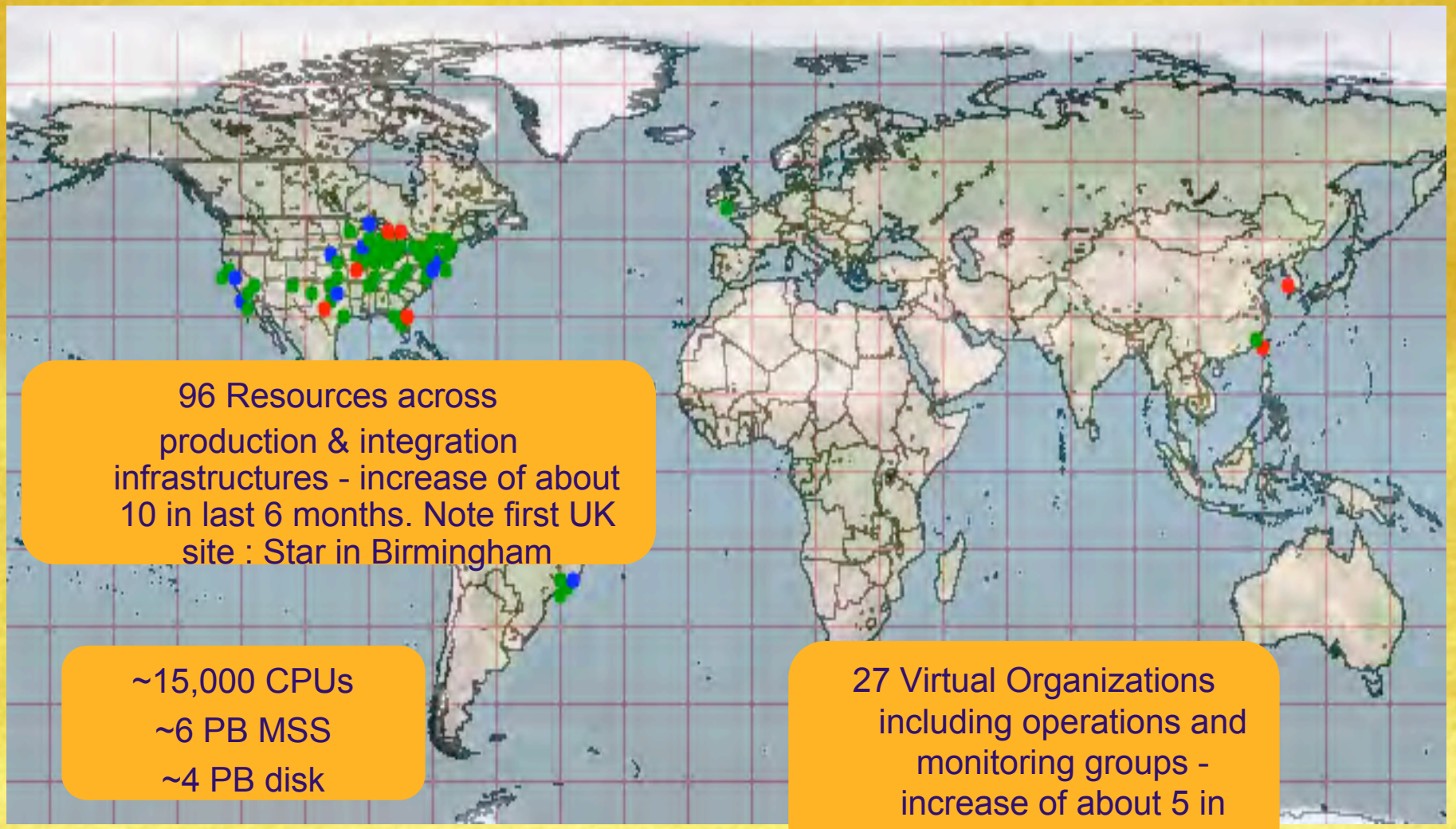
Who is OSG ?

Open Science Grid

- Many DOE Labs and DOE/NSF sponsored University IT facilities.
- Grid technology groups: Condor, Globus, SRM, NMI.
- Large global physics collaborations: US ATLAS, US CMS, LIGO, CDF, D0, STAR
- Smaller research collaborations: SDSS, DES, GADU,
- Partnerships e.g. TeraGrid, European Grids, Regional/Campus Grids e.g. Texas, Wisconsin…
- Education projects e.g. I2U2.

Current OSG deployment

Open Science Grid



96 Resources across production & integration infrastructures - increase of about 10 in last 6 months. Note first UK site : Star in Birmingham

~15,000 CPUs
~6 PB MSS
~4 PB disk

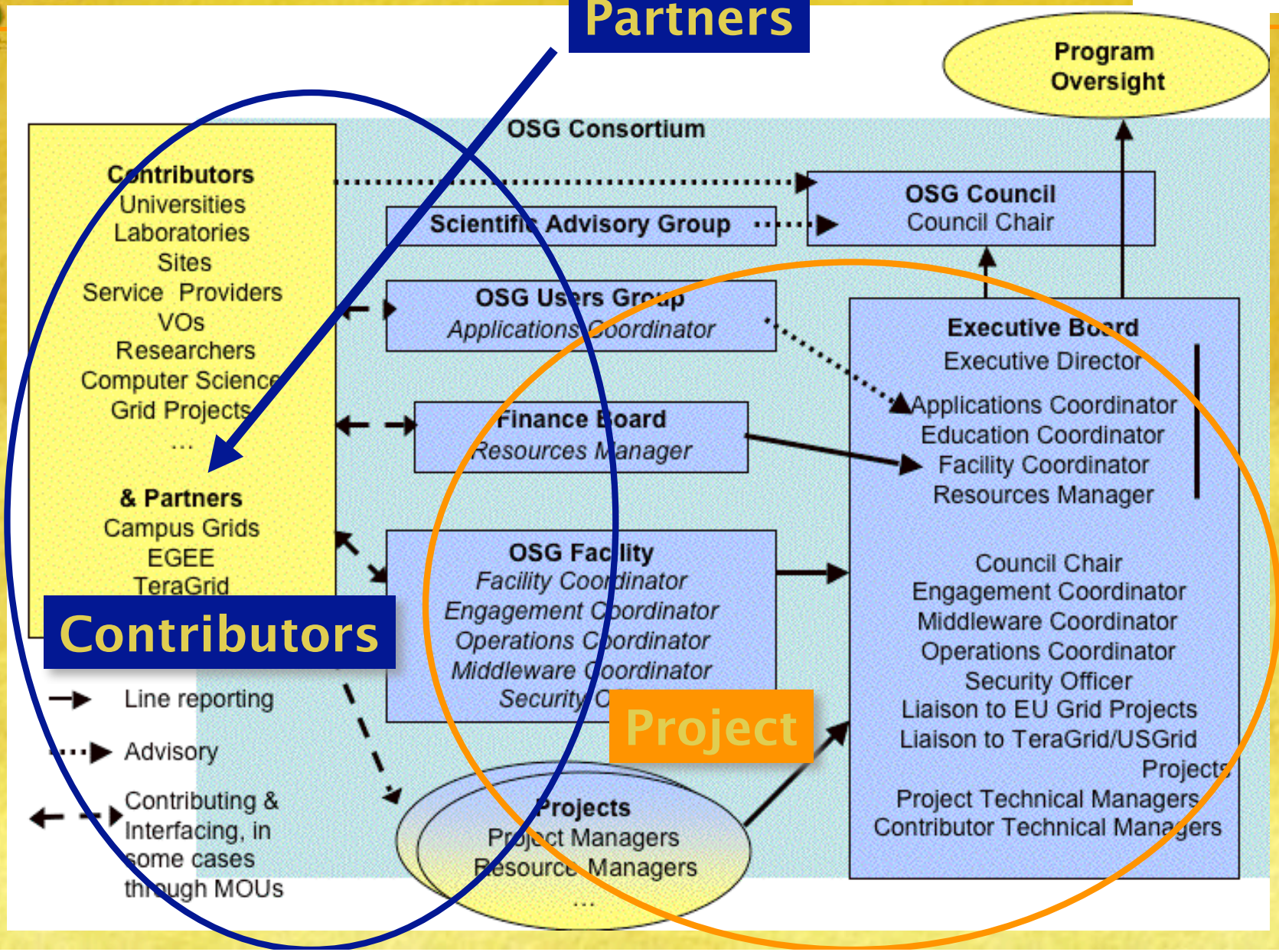
27 Virtual Organizations including operations and monitoring groups - increase of about 5 in last 6 months

OSG Funding

- OSG will be co-funded by DOE and NSF for 5 years at $6M/year starting in Sept ‘06
- Includes major deliverables for US LHC, LIGO and STAR; well used by CDF and D0;
- Commitment to engage communities beyond Physics.
- Commitment to collaborate with EGEE and TeraGrid.
- Consortium policies are to be open to participation by all researchers.
- Project responsibilities are to operate, protect, extend and support the Distributed Facility for the Consortium.

OSG Consortium

Open Science Grid



OSG & Outside-of-physics Communities

Open Science Grid

- **Genome analysis GADU** project re-analyses publicly available genome databases and makes usable, annotated, summaries for researchers
- >100 **Nanotechnology Biomoca** jobs -- that run from 10-40 days -- are being executed on LIGO, ATLAS and CMS sites.
- We are discussing partnership with the Northwest Indiana Computing Grid (NWICG) -- which would bring in **Gaussian.**
- Alan Blatecky's group at RENCI is porting the "award winning **Bioportal**" to OSG.
- The **P-Grade portal** running a version of **CHARMM** molecular dynamics simulation package.
- Work on **Campus Grids** enabling Crimson Grid, NWICG, New York State Grid (NYSG) , GPN (Nebraska Education/Training grid) partnerships. (Note: Partners do not have to contribute resources; collaboration can equally be in software, procedures, education, training, security etc.)

OSG Metrics - Sept 2006

Open Science Grid

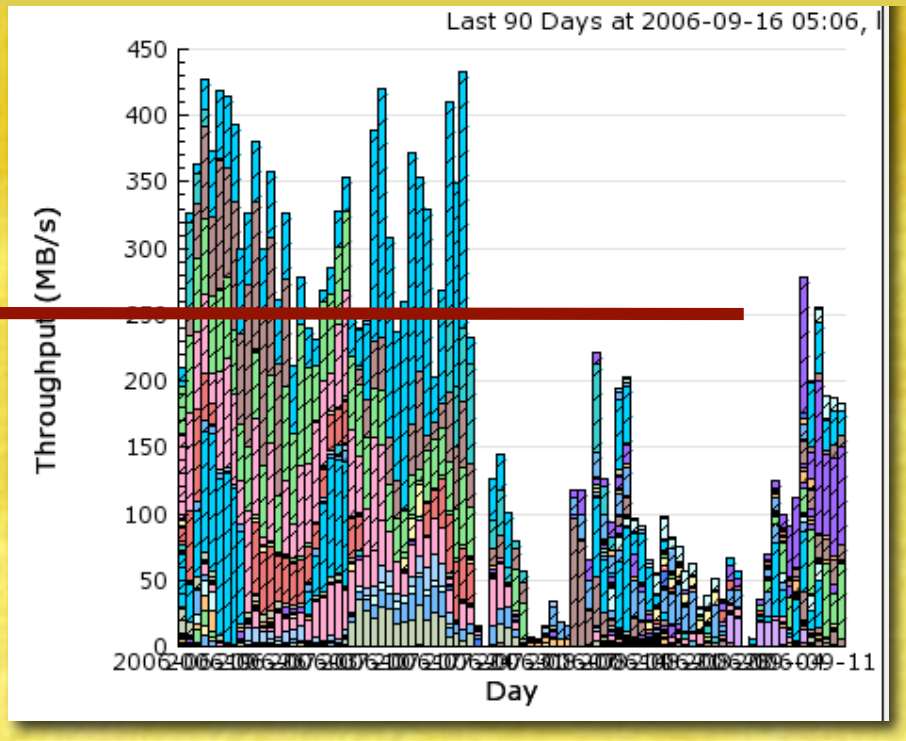
Compute sites	59
Storage sites	11 (3 have tape and 8 are disk only)
Accessible CPUs (excluding those in partner Campus Grids)	15K Commodity linux boxes
Custodial Storage	6 Petabytes (owned by specific VOs and in general non-sharable)
Disk storage and caches	2 PB
Data transfer (include from CERN to Tier-1s and Tier-1s to Tier-2s)	~2-5Gbits/sec
Number of registered VOs (including operations/monitoring)	28 (4 Education projects + 7 Outside-of-physics.)
Number of simultaneiously running jobs	3K-4K
CPUhours per day	50K

High Throughput Data Transfer

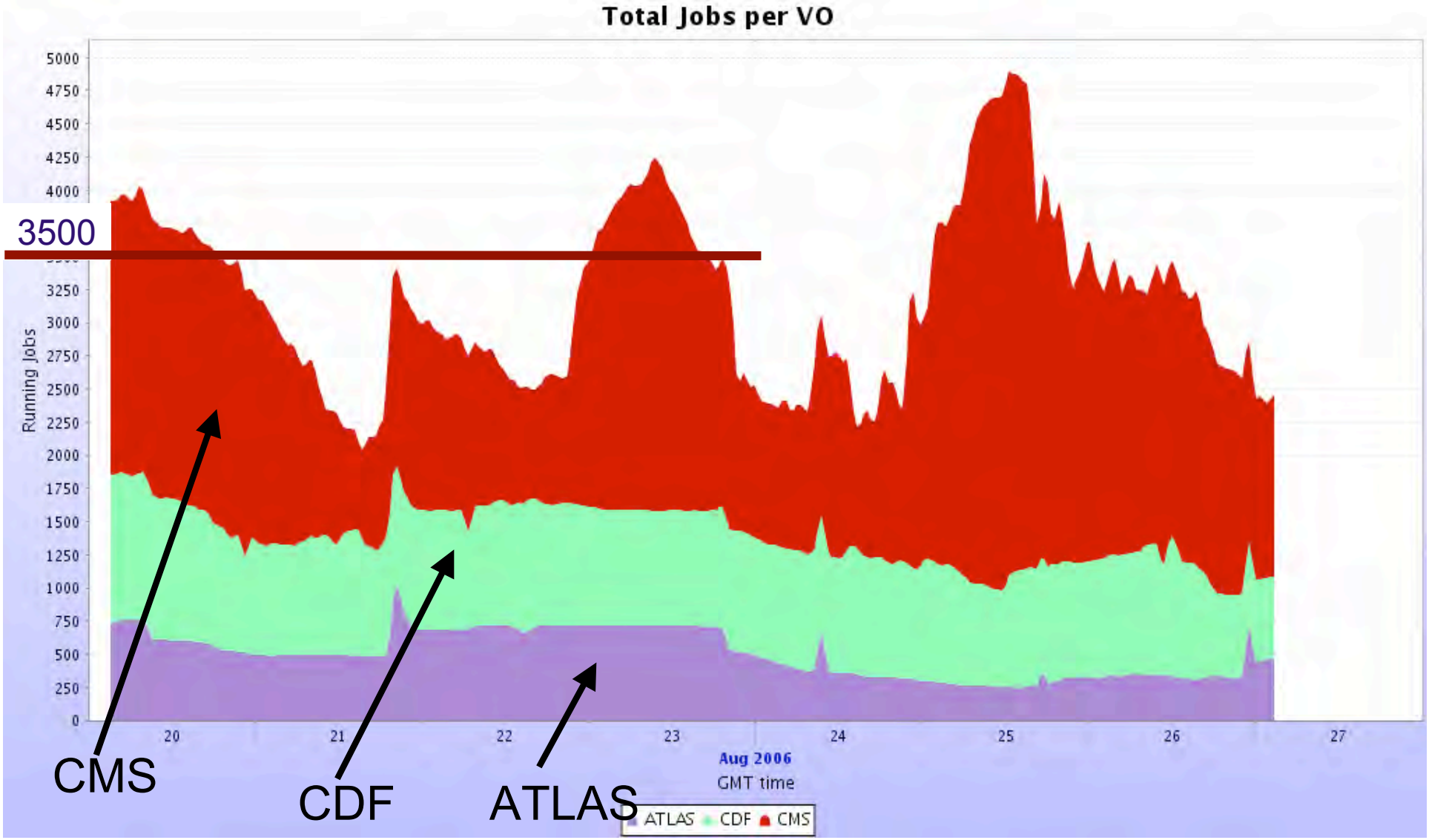
- OSG must deliver petabytes of data moved and stored (and removed) for the US LHC in 2008.
- LIGO data scales will increase (factor of 10?) in 2009.

Last 3 months throughput to US CMS Tier-2s:

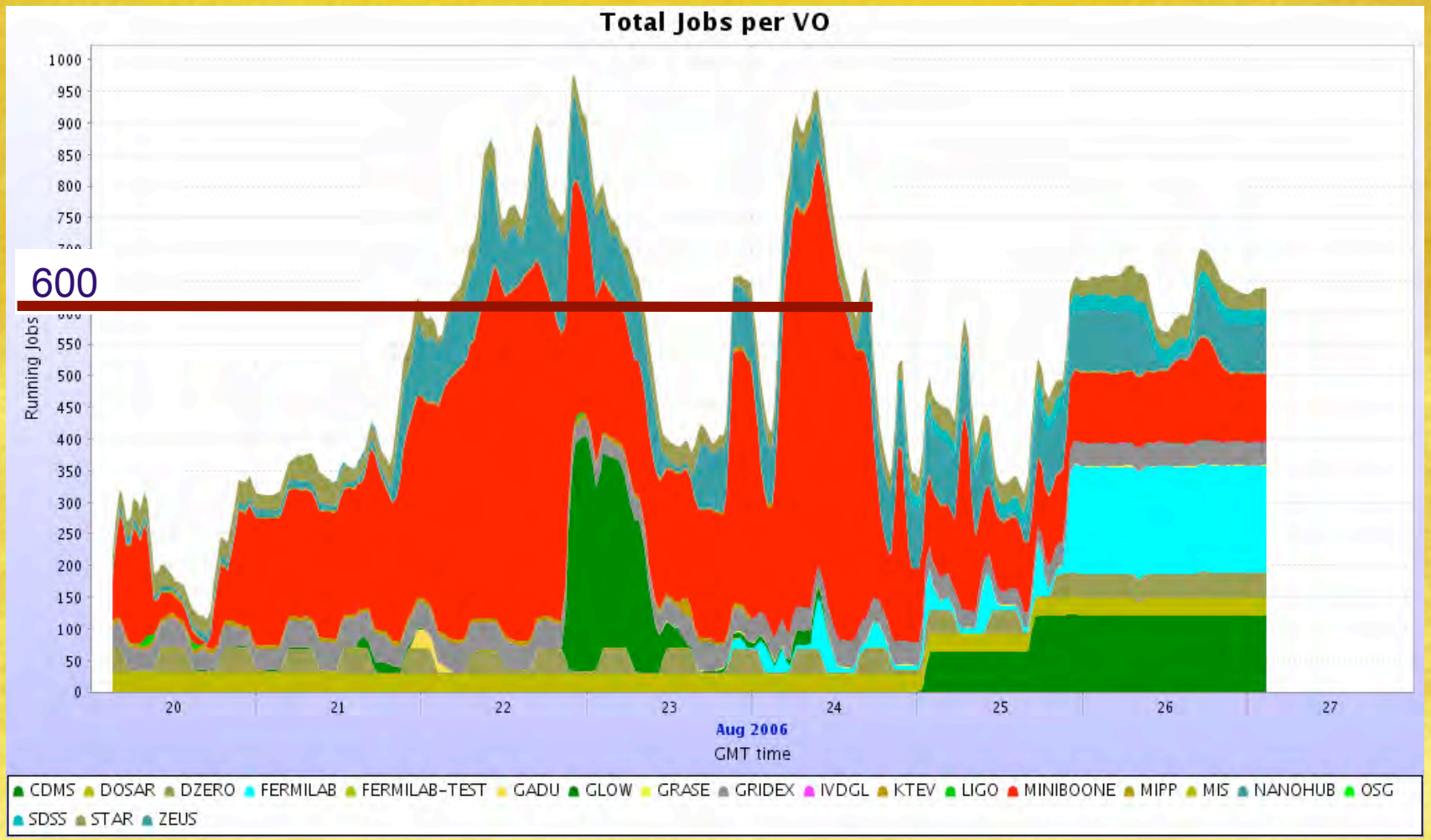
1 GBit/sec



The big three run continuously



Mix of smaller VOs..



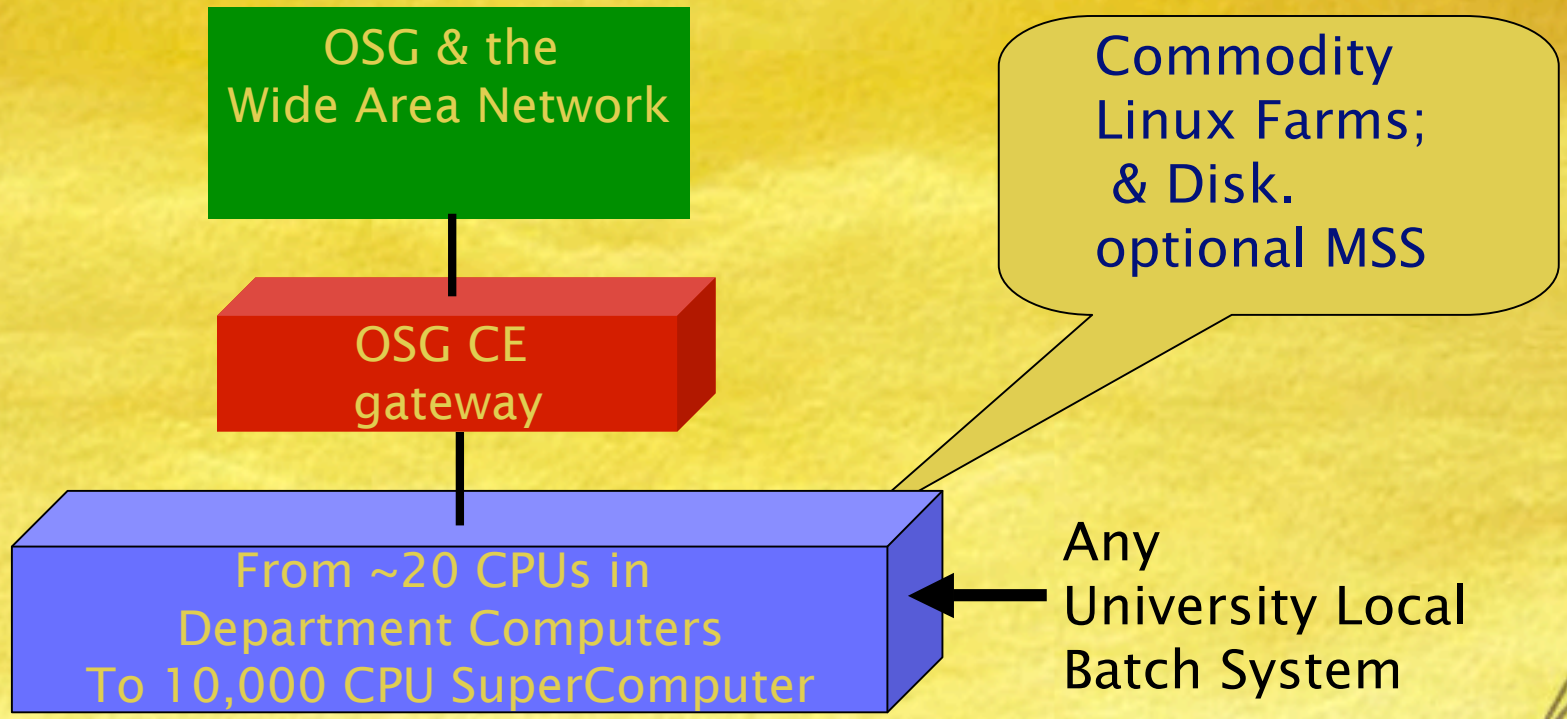
What is the OSG?

Compute Elements

Open Science Grid

Processing Farms accessed through GRAM and supporting job execution through one of 5 different batch systems.

Priorities and policies through assignment of Group/VO Roles mapped to accounts and batch queue priorities.

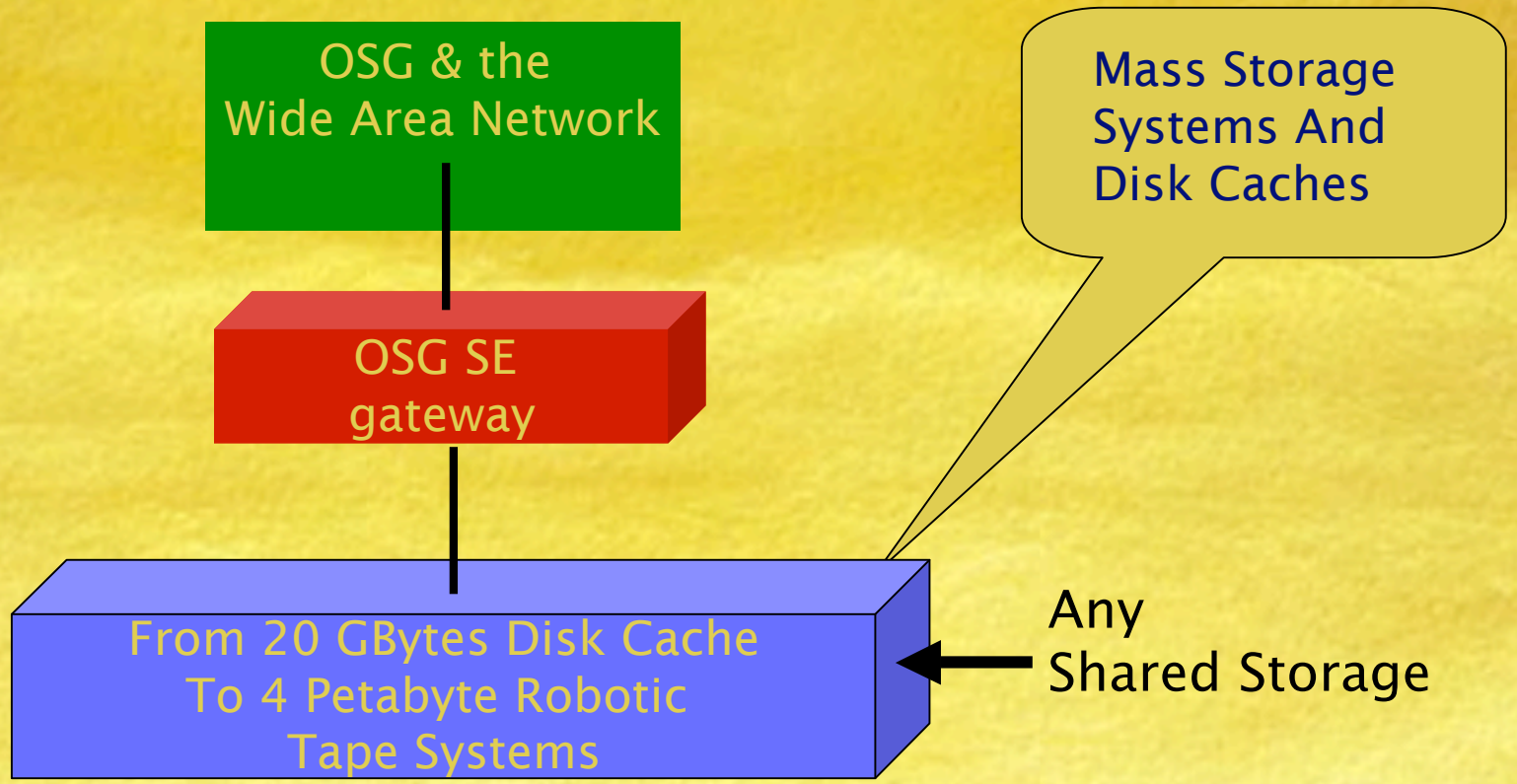


Storage Elements

Open Science Grid

Storage Services - access storage through Storage Resource Manager (SRM) interface and GridFtp.

Allocation of shared storage through agreements between Site and VO(s) facilitated by OSG.



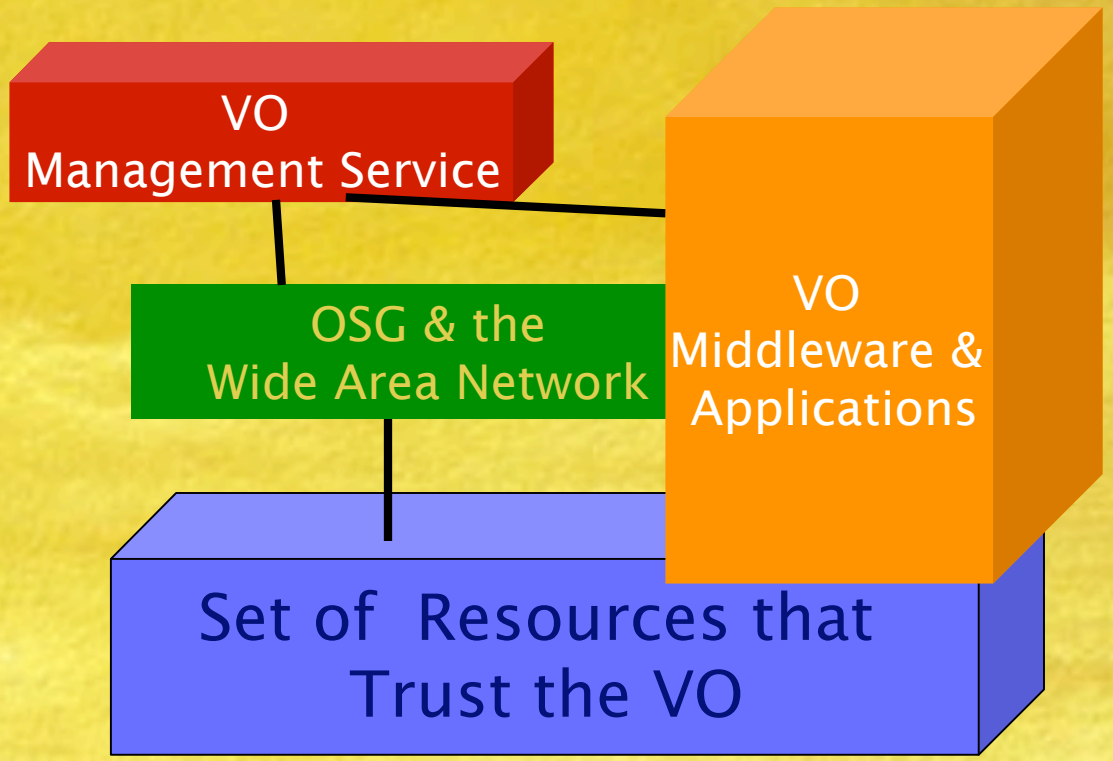
Virtual Organizations (VOs)

OSG Infrastructure trades in Groups not Individuals.

Virtual Organization Management services allow registration, administration and control of members of the group.

Facilities trust and authorize VOs.

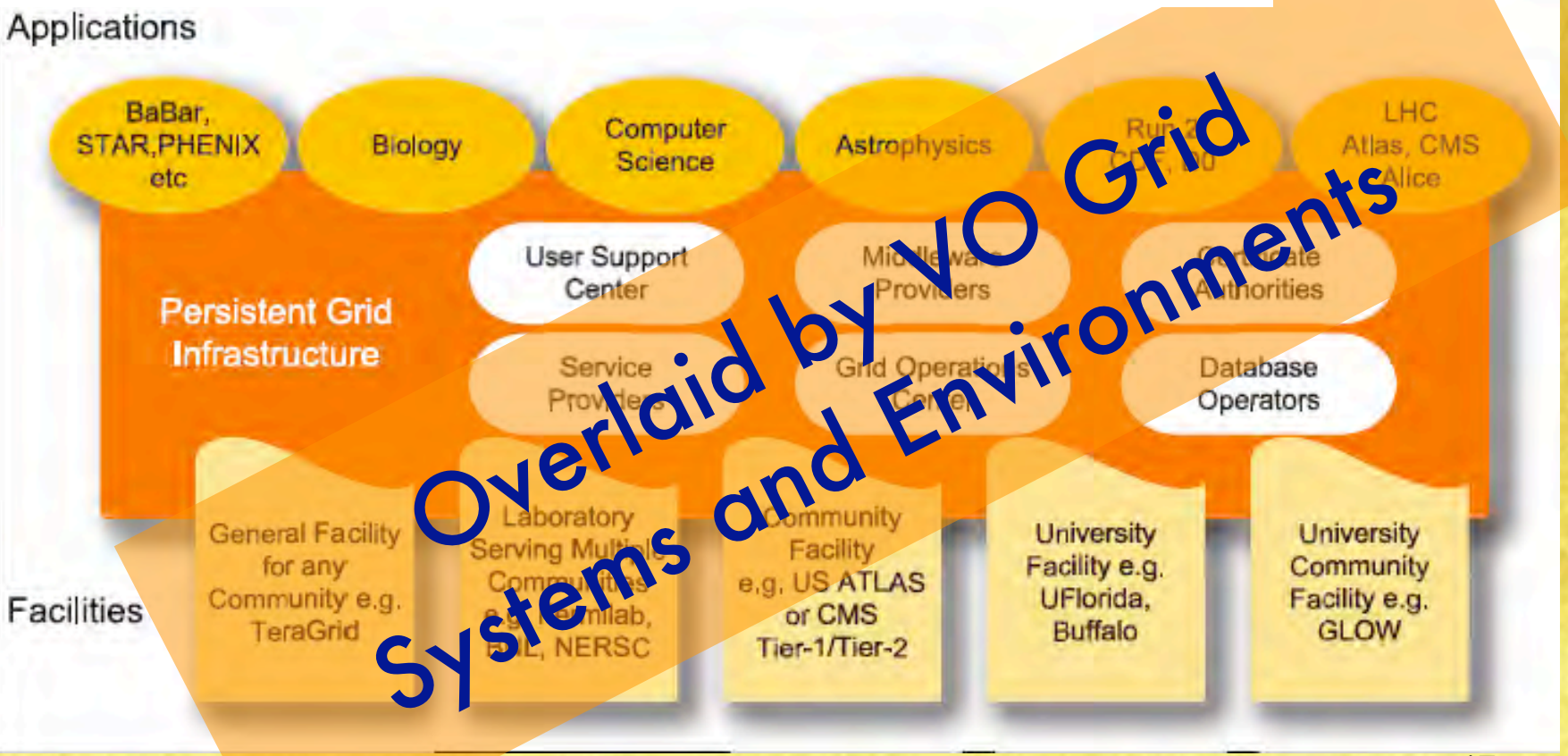
Storage and Compute Services prioritize according to VO group.



Common Grid Infrastructure

Open Science Grid

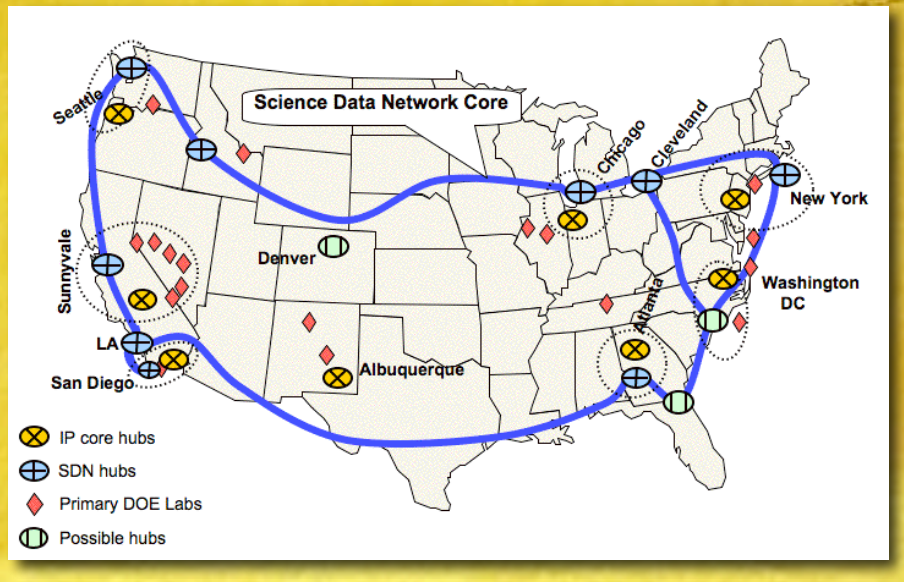
OSG provides gridwide monitoring, information, validation, accounting services.

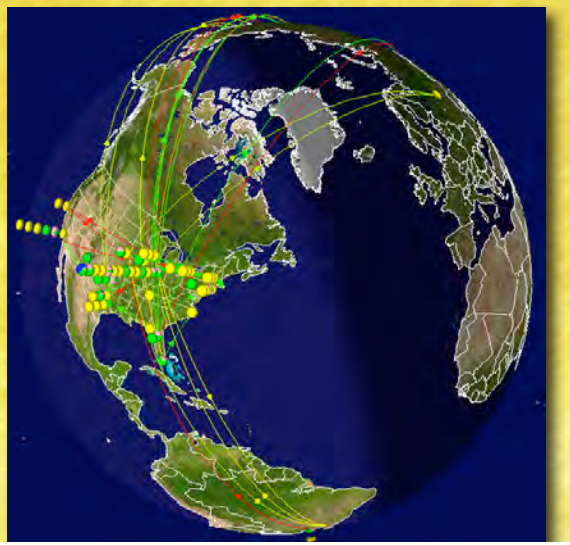


Network Connectivity

Open Science Grid

- Use commodity networks - ESNet, Campus LANs
- Well network provisioned sites e.g. connected to Starlight to low bandwidth connections e.g. Taiwan
- Connectivity ranges from full-duplex, outgoing only, to fully behind firewalls.
- Expect more focus on network issues as data aspects become more important.





Middleware

OSG Middleware

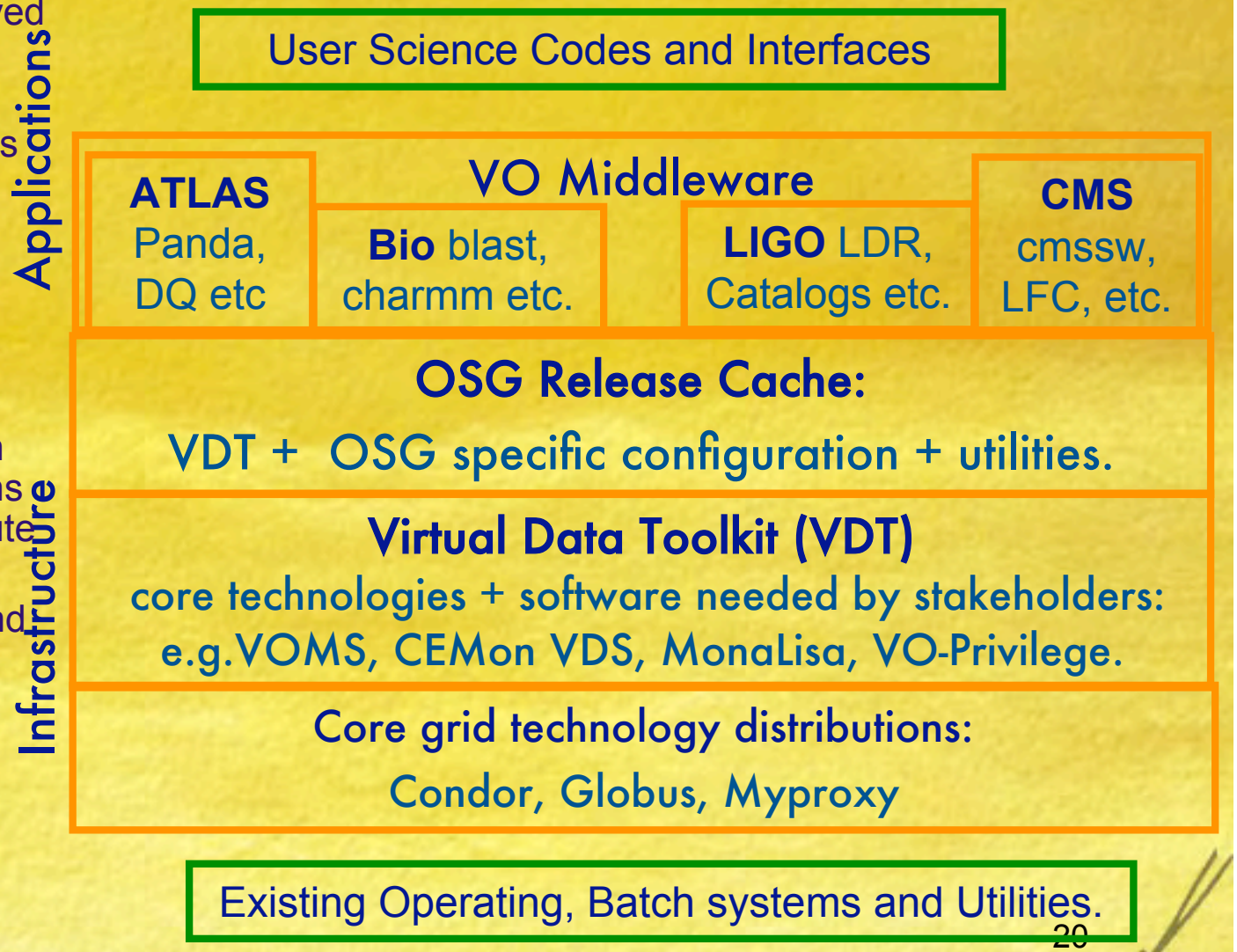
Open Science Grid

OSG Middleware is deployed on existing farms and storage systems.

OSG Middleware interfaces to the existing installations of OS, utilities and batch systems.

VOs have VO scoped environments in which they deploy applications (and other files), execute code and store data.

VOs are responsible for and have control over their end-to-end distributed system using the OSG infrastructure.

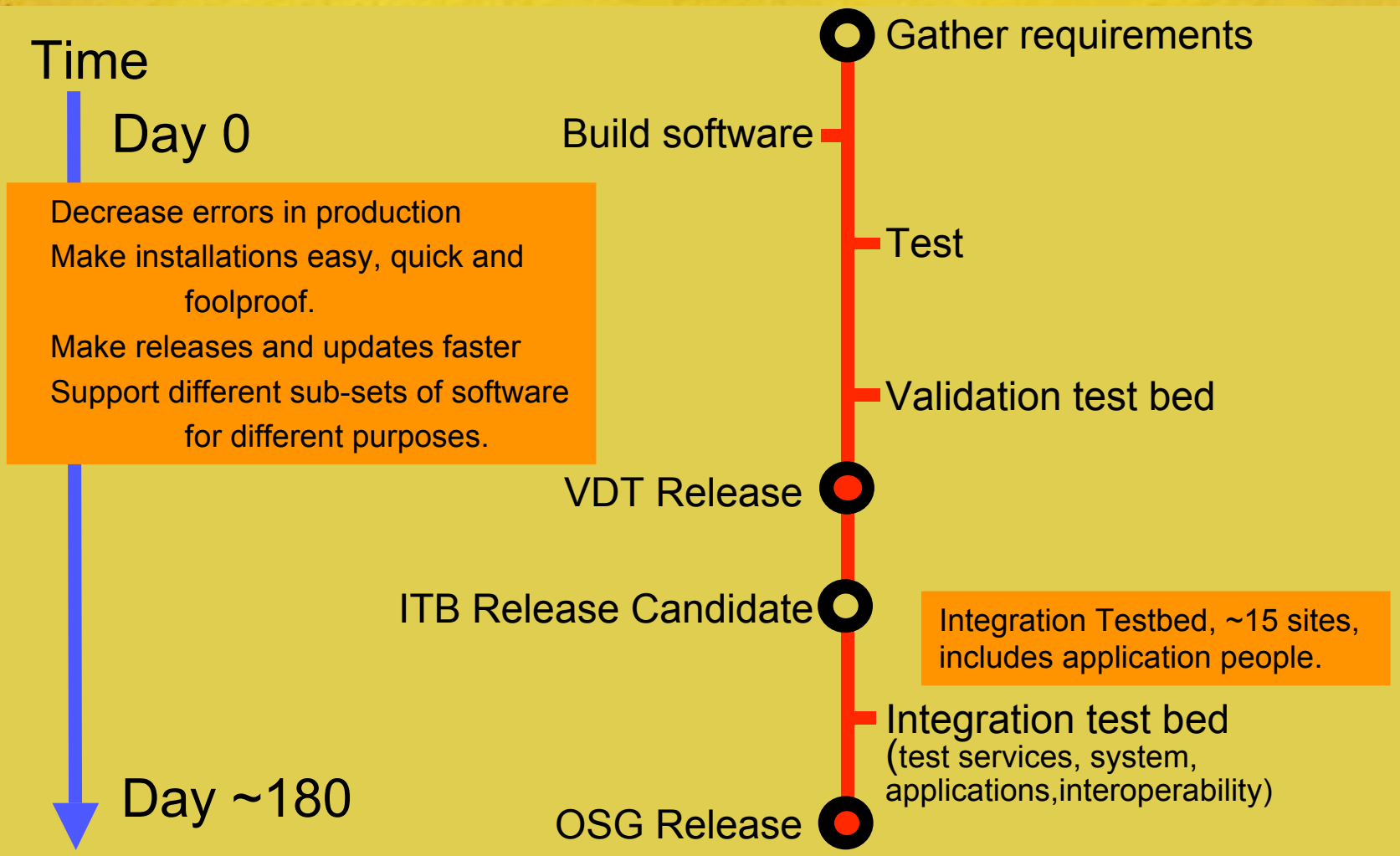


What is the VDT?

Open Science Grid

- A collection of software
  - Grid software (Condor, Globus and lots more)
  - Virtual Data System (Origin of the name “VDT”)
  - Utilities
  - Built for >10 flavors/versions of Linux
- An easy installation
  - Goal: Push a button, everything just works
  - Two methods:
    - Pacman: installs and configures it all
    - RPM: installs some of the software, no configuration
- A support infrastructure

Software Release Process



What software is in the VDT?

Open Science Grid

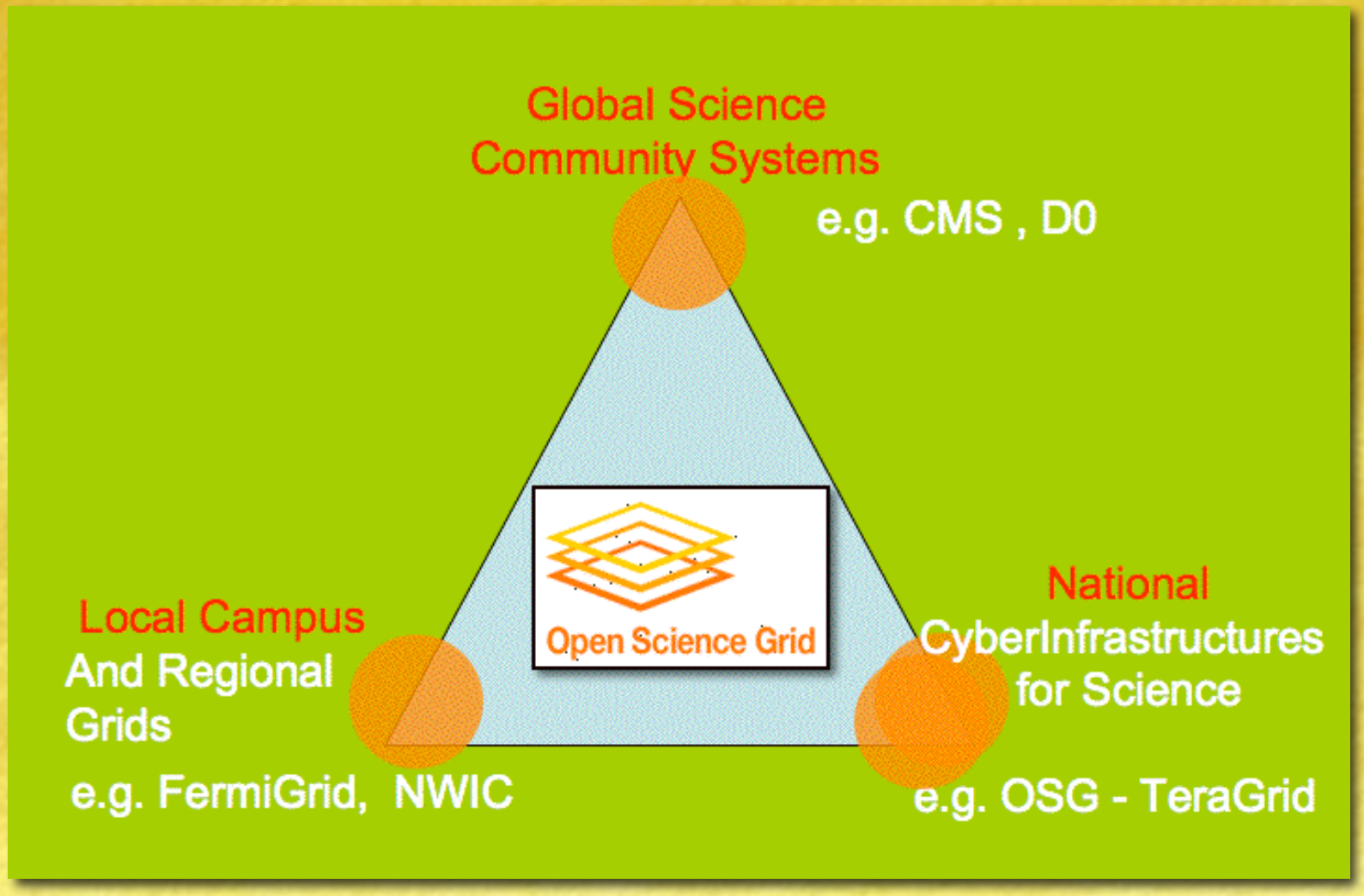
- Job Management
  - Condor (& Condor-G & Condor-C)
  - Globus GRAM
- Data Management
  - GridFTP (data transfer)
  - RLS (replication location)
  - DRM (storage management)
  - Globus RFT
- Information Services
  - Globus MDS
  - GLUE schema & providers
- Client tools
  - Virtual Data System
  - SRM clients (V1 and V2)
  - UberFTP (GridFTP client)
- Developer Tools
  - PyGlobus, PyGridWare
- Testing
  - NMI Build & Test, VDT Tests
- Security
  - VOMS (VO membership)
  - GUMS (local authorization)
  - mkgridmap (local authorization)
  - MyProxy (proxy management)
  - GSI SSH
  - CA CRL updater
- Monitoring
  - MonaLISA
  - gLite CEMon
- Accounting
  - OSG Gratia
- Support
  - Apache,Tomcat
  - MySQL (with MyODBC)
  - Non-standard Perl modules
  - Wget
  - Squid
  - Logrotate, Configuration Scripts

Due diligence to Security

- Risk assessment, planning, Service auditing and checking
- Incident response, Awareness and Training, Configuration management,
- User access Authentication and Revocation, Auditing and analysis. End to end *trust* in quality of code executed on remote CPU - signatures?
- Identity and Authorization: Extended X509 Certificates
  - OSG is a founding member of the US TAGPMA.
  - DOEGrids provides script utilities for bulk requests of Host certs, CRL checking etc.
  - VOMS extended attributes and infrastructure for Role Based Access Controls.

One Grid Among Many

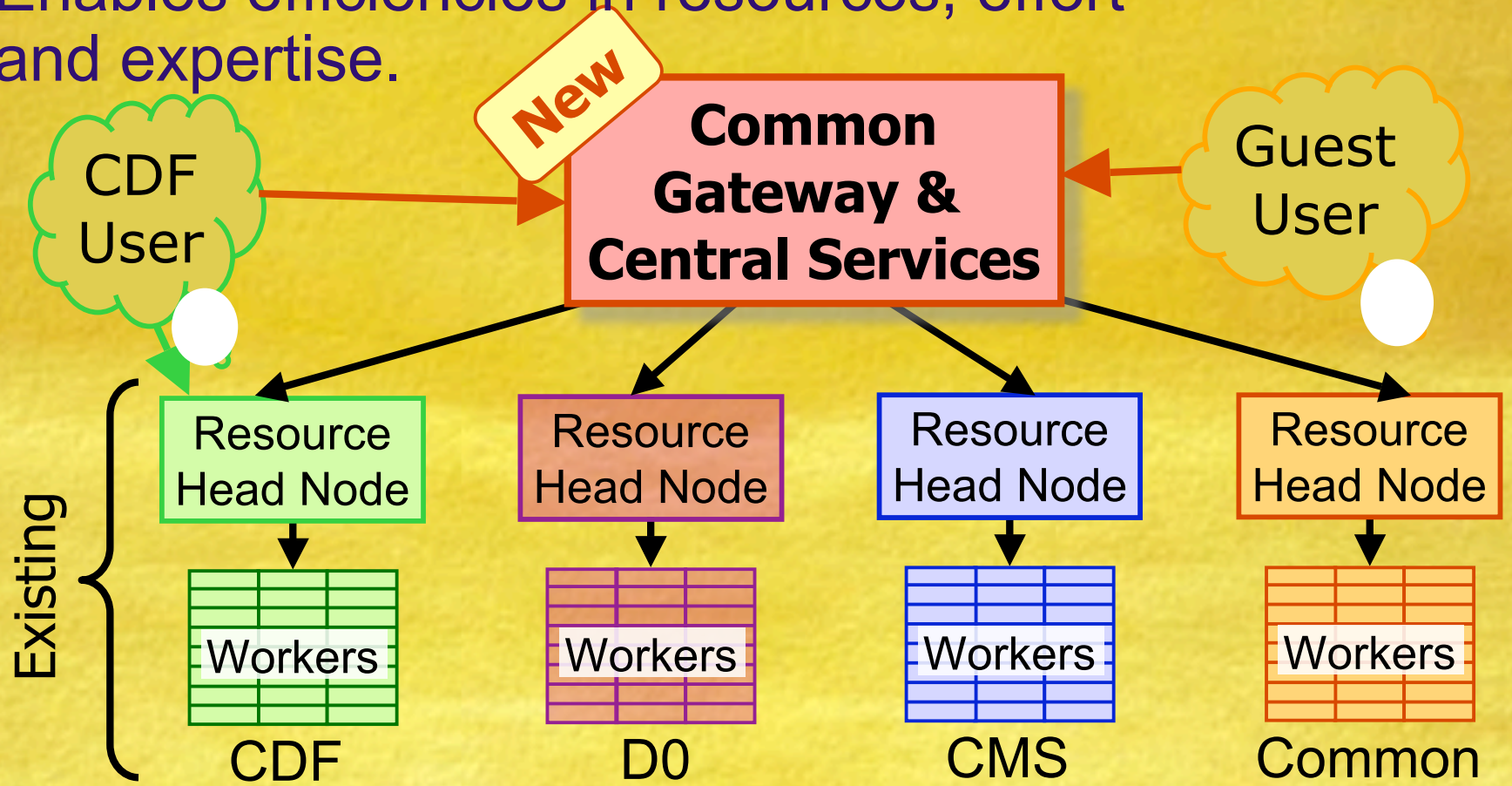
Grid - of - Grids



FermiGrid Exemplar Campus Grid

Open Science Grid

- Enables efficiencies in resources, effort and expertise.



New Resources to be added to "Common"

Bridging Campus Grid to OSG: GLOW

Open Science Grid

Dispatch jobs from local security, job, storage infrastructure and "uploading" to wide-area infrastructure.

Communication

www.opensciencegrid.org
http://vdt.cs.wisc.edu
http://osg.ivdgl.org/twiki/bin/view

Training - e.g. Grid Summer Workshop Year 4

Open Science Grid

SummerGridSyllabus2006 < SummerGridWorkshop < TWiki 08/10/2006 04:24 PM

TWiki > SummerGridWorkshop Web > SummerGridSyllabus2006 r31 - 29 Jun 2006 - 15:48:24 - BenClifford

Summer Grid Workshop 2006 Syllabus

Greetings

- Welcome to UT Brownsville and the Workshop: Dr. Manuela Campanelli, Center for Gravita Astronomy and Department of Physics, University of Texas, Brownsville.
- Welcome - from Ruth Pordes, Executive Director, Open Science Grid
- Workshop overview - Mike Wilde, Director of Education, Outreach and Training, Open Scie

Lecture 1 - Monday AM: Grid Intro and Fundamentals Review

- **Topics:** Networking concepts review, Clusters, Grids, e-Science, fundamental Grid compon Technologies

Hands on.

Technical trainers. Nice Setting (Padre Island).

Students got their own applications to run on OSG!

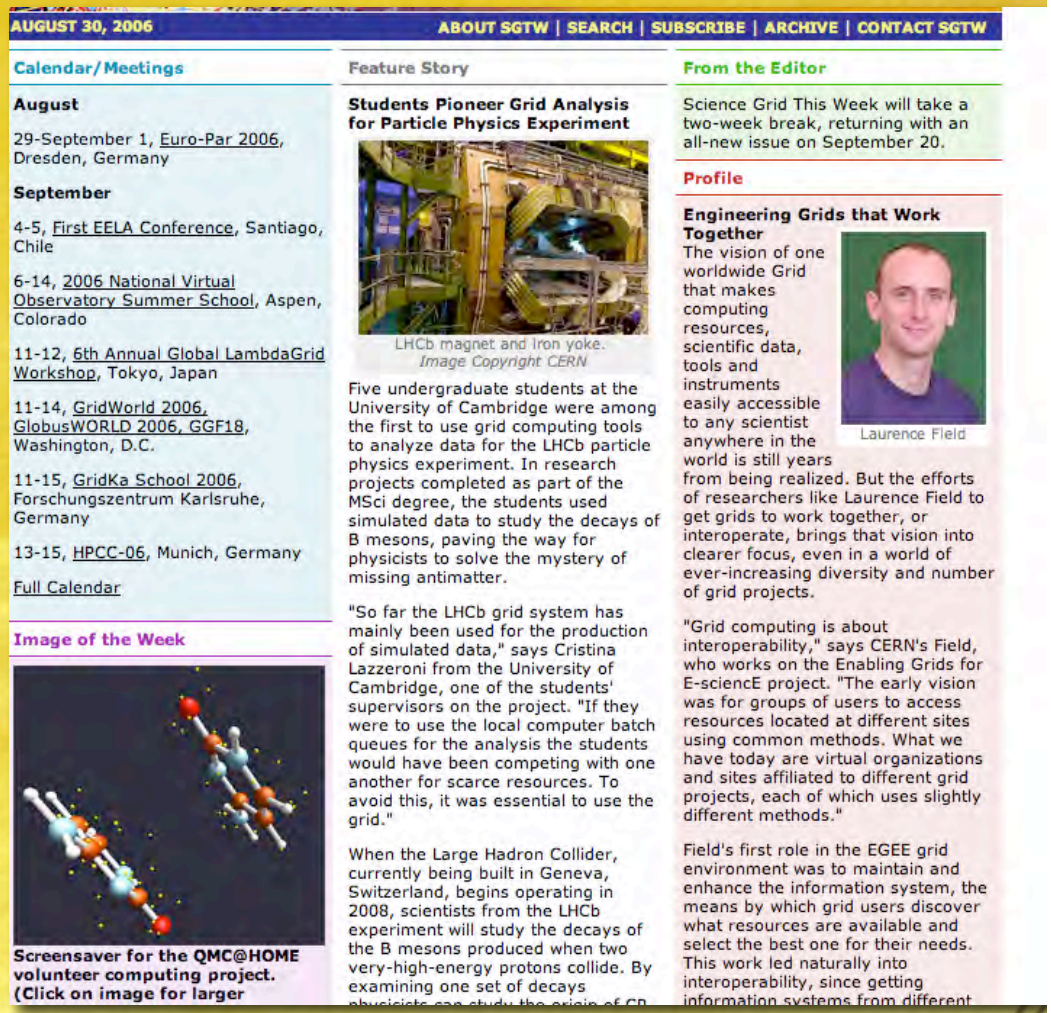
Science Grid This Week

Open Science Grid

Broad set of activities

- (Katie Yurkewicz)
- News releases, PR, etc.
- Science Grid This Week, 1.5 years (>1000 subscribers)

www.interactions.org/sgtw



AUGUST 30, 2006 | ABOUT SGTW | SEARCH | SUBSCRIBE | ARCHIVE | CONTACT SGTW

Calendar/Meetings

August

29-September 1, Euro-Par 2006, Dresden, Germany

September

4-5, First EELA Conference, Santiago, Chile

6-14, 2006 National Virtual Observatory Summer School, Aspen, Colorado

11-12, 6th Annual Global LambdaGrid Workshop, Tokyo, Japan

11-14, GridWorld 2006, GlobusWORLD 2006, GGF18, Washington, D.C.

11-15, GridKa School 2006, Forschungszentrum Karlsruhe, Germany

13-15, HPCC-06, Munich, Germany

Full Calendar

Image of the Week

Screensaver for the QMC@HOME volunteer computing project. (Click on image for larger

Feature Story

Students Pioneer Grid Analysis for Particle Physics Experiment

LHCb magnet and iron yoke. *Image Copyright CERN*

Five undergraduate students at the University of Cambridge were among the first to use grid computing tools to analyze data for the LHCb particle physics experiment. In research projects completed as part of the MSci degree, the students used simulated data to study the decays of B mesons, paving the way for physicists to solve the mystery of missing antimatter.

"So far the LHCb grid system has mainly been used for the production of simulated data," says Cristina Lazzeroni from the University of Cambridge, one of the students' supervisors on the project. "If they were to use the local computer batch queues for the analysis the students would have been competing with one another for scarce resources. To avoid this, it was essential to use the grid."

When the Large Hadron Collider, currently being built in Geneva, Switzerland, begins operating in 2008, scientists from the LHCb experiment will study the decays of the B mesons produced when two very-high-energy protons collide. By examining one set of decays

From the Editor

Science Grid This Week will take a two-week break, returning with an all-new issue on September 20.

Profile

Engineering Grids that Work Together

Laurence Field

The vision of one worldwide Grid that makes computing resources, scientific data, tools and instruments easily accessible to any scientist anywhere in the world is still years from being realized. But the efforts of researchers like Laurence Field to get grids to work together, or interoperate, brings that vision into clearer focus, even in a world of ever-increasing diversity and number of grid projects.

"Grid computing is about interoperability," says CERN's Field, who works on the Enabling Grids for E-sciencE project. "The early vision was for groups of users to access resources located at different sites using common methods. What we have today are virtual organizations and sites affiliated to different grid projects, each of which uses slightly different methods."

Field's first role in the EGEE grid environment was to maintain and enhance the information system, the means by which grid users discover what resources are available and select the best one for their needs. This work led naturally into interoperability, since getting information systems from different

Summary

Open Science Grid

- OSG provides some core services and software, provides a distributed facility for its stakeholders, and partners with other projects to share experience and collaborate in several different areas.